

# BAMO at SemEval-2024 Task 9: BRAINTEASER: A Novel Task Defying Common Sense

Baktash Ansari, Mohammadmostafa Rostamkhani, Sauleh Eetemadi

Iran University of Science and Technology

{baktash\_ansari, mo\_rostamkhani97}@comp.iust.ac.ir, sauleh@iust.ac.ir

## Abstract

This paper outlines our approach to SemEval 2024 Task 9, BRAINTEASER: A Novel Task Defying Common Sense. The task aims to evaluate the ability of language models to think creatively. The dataset comprises multi-choice questions that challenge models to think 'outside of the box'. We fine-tune 2 models, BERT and RoBERTa Large. Next, we employ a Chain of Thought (CoT) zero-shot prompting approach with 6 large language models, such as GPT-3.5, Mixtral, and Llama2. Finally, we utilize ReConcile, a technique that employs a 'round table conference' approach with multiple agents for zero-shot learning, to generate consensus answers among 3 selected language models. Our best method achieves an overall accuracy of 85 percent on the sentence puzzles subtask.

## 1 Introduction

Evaluation methods in the NLP community predominantly emphasize Vertical thinking, characterized by sequential, analytical processes based on rationality, logic, and rules. However, SemEval-2024 Task 9, BRAINTEASER (Jiang et al., 2024b), which is based on the original BRAINTEASER dataset (Jiang et al., 2023), aims to introduce a task that promotes lateral thinking (or "thinking outside the box"), a divergent and creative process involving the exploration of new perspectives when addressing problems. The BRAINTEASER QA task consists of two subtasks for the English language: Sentence Puzzles and Word Puzzles. This task is designed to challenge the common sense reasoning capabilities of NLP models and stimulate the development of models that can think laterally.

- **Sentence Puzzles:** Sentence-type brain teaser where the puzzle-defying commonsense is centered on sentence snippets.
- **Word Puzzles:** Word-type brain teaser where the answer violates the default meaning of the

word and focuses on the letter composition of the target question.

We generate baselines for two attention-based models, BERT (Devlin et al., 2019) and RoBERTa-Large (Liu et al., 2019), as selected in the task paper, to solve these types of multiple-choice problems. Then we fine-tune them with the same configs. After achieving some accuracy through fine-tuning, we explore zero-shot prompting with various large language models (LLMs). To further improve results, zero-shot prompting is conducted using a Chain of Thought technique (Wei et al., 2023). As illustrated in Figure 1, we compel the model to analyze and provide step-by-step reasoning for its answer instead of simply providing a correct option alone. This approach helps the model focus more on details and answer questions with fewer errors.

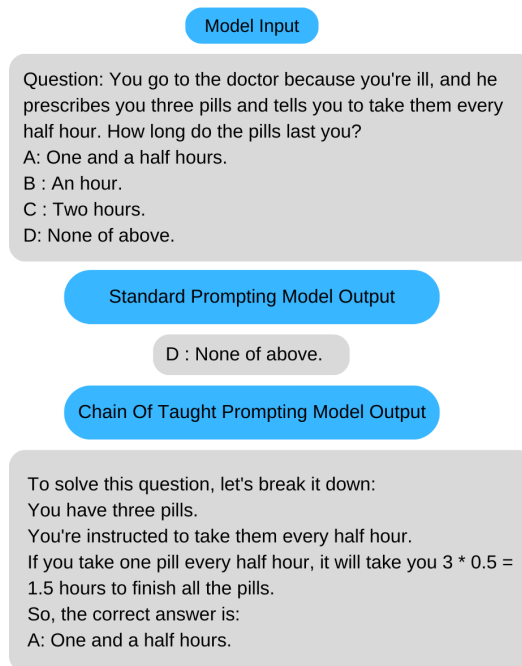


Figure 1: Chain Of Thought Prompting (GPT3.5)

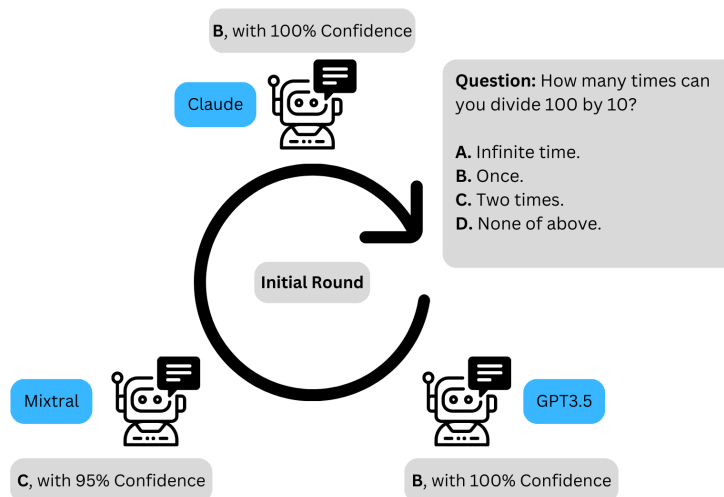


Figure 2: An Illustration of RECONCILE for Initial Round

Communication among multiple agents is fundamentally important in complex decision-making processes. Therefore, as our main strategy, we use the ReConcile technique (Chen et al., 2023), which generates answers by gathering the consensus of multiple models based on their confidence levels, to address these problems. To make this technique compatible with this type of dataset, we extend the application of ReConcile from yes/no questions to the BRAINTEASER questions. In this system, we apply the concept of a society of minds to multiple agents. For round table decision-making, we execute three phases: Initial Response Generation, Multi-Round Discussion, and Final Answer Generation on three language learning models: Mixtral8x7b (Jiang et al., 2024a), Claude<sup>1</sup>, and GPT3.5<sup>2</sup>. As illustrated in Figure 2, in each phase, we generate specific prompts (see Appendix B) for models to answer the question, along with their confidence level between 0 and 1. We then use their answers for the next round and derive an overall consensus answer for each round. The method that gave us our best result achieved a rank of 11 out of 33 for the sentence puzzles in the task leaderboard. Further details of our implementations are available through our GitHub repository.<sup>3</sup>

## 2 Background

### 2.1 Related Works

The exploration of reasoning abilities in large language models, lateral thinking, and common sense

reasoning has been the focus of several studies in recent years. The BRAINTEASER is a novel task in this context, requiring a unique blend of these capabilities. In the paper (Zhang et al., 2022), foundational insights into the use of knowledge graphs for self-supervision in common sense reasoning tasks are provided. This work is particularly relevant as it demonstrates how external knowledge can significantly improve an LLM’s ability to solve reasoning problems. Furthering the discussion on reasoning, LatEval (Huang et al., 2024) introduces an evaluation benchmark for LLMs based on lateral thinking puzzles. This process is similar to the method in solving the BRAINTEASER, which encourages models to elaborate on their reasoning, leading to improved performance. The paper RiddleSense (Lin et al., 2021) offers an analysis of how LLMs handle riddle questions that require linguistic creativity and common sense knowledge. This is directly linked to our approaches, where the ability of an LLM to use analogous examples influences its problem-solving effectiveness. Also, (Dou and Peng, 2022) investigates the inherent knowledge within language models and their ability to function in zero-shot common sense question answering tasks. The integration of external knowledge is particularly relevant to tasks requiring common sense and real-world information, underscoring the importance of knowledge retrieval in the context of the BRAINTEASER. MVP-Tuning (Huang et al., 2023) introduces a novel approach to knowledge retrieval using prompt tuning. This aligns with our aim in solving BRAINTEASER to evaluate the capacity of LLMs to handle questions that necessitate an understanding of the world as humans

<sup>1</sup>Available at <https://claude.ai/>.

<sup>2</sup>Available at <https://openai.com/>.

<sup>3</sup>GitHub Repository

perceive it. Lastly, ReConcile (Chen et al., 2023) and (Liang et al., 2023) both discuss the use of multiple LLMs to improve reasoning capabilities. These works highlight the potential of using a diverse set of models to solve complex tasks like the BRAINTEASER, further enriching the strategies for solving such tasks with AI systems.

## 2.2 Datasets

The organizers provide datasets for one language: English. As mentioned previously, the dataset consists of two categories: Sentence Puzzles and Word Puzzles. The task providers construct reconstruction versions of the original data in two parallel ways: Semantic Reconstruction and Context Reconstruction. This is done to ensure that the task evaluates lateral thinking ability rather than mere memorization. The Semantic variant reformulates the initial question while preserving its answer, whereas the Context variant retains the misleading commonsense assumption as is and modifies the question and its answer to fit a different situational context.

The dataset is split into two parts for the evaluation phase: train and test sets for each category. In the sentence puzzles category, the train set comprises 507-row samples, while the test set consists of 120-row samples. Similarly, in the word puzzles category, the train set contains 396-row samples, and the test set has 96-row samples. Each sample includes a question with its corresponding answer and three distractors. We utilize both word and sentence puzzle datasets during the training phase, but only the sentence puzzles dataset is used for zero-shot phases.

## 2.3 Evaluation Metrics

The accuracy metric is employed for evaluation as described in the task paper. Performance evaluation is conducted using two accuracy metrics: Instance-based Accuracy and Group-based Accuracy for Original, Semantic, and Context questions.

# 3 System overview

## 3.1 Preprocessing

In the fine-tuning phase, we employ two transformer-based models, BERT and RoBERTa, for multiple choice tasks. Both models are pre-trained on large text corpora. The input to these models is a sequence constructed by concatenating the question with each choice, separated by special

tokens. This process is facilitated by the models’ tokenizers, which convert the text into a format that the models can understand. For a given question "Q", and choices, the input to the models would be:

$$\text{Input}_i = [\text{CLS}] \text{ Q } [\text{SEP}] \text{Choice}_i [\text{SEP}] \quad (1)$$

where "i" represents the index of the choice. Each sequence represents a different choice, and the models’ understanding of the context and the choices allows them to predict the correct answer. The fine-tuning process adapts the models to this specific task, optimizing their parameters to minimize the difference between the predicted and actual answers.

## 3.2 Model Training

We utilize BERT-Base and RoBERTa-Large models, along with their respective tokenizers, for word embeddings in the multiple choice task, sourced from the Hugging Face library. These models are finetuned using the Hugging Face trainer. Initially, we load the models and establish a baseline on the test set for both sentence and word puzzles. Subsequently, each model undergoes finetuning with the default Hugging Face Cross-Entropy loss function for classification, and overall accuracy is computed as the metric using the same hyper-parameter configurations as detailed in Table 1.

Hyperparameter	Value
Learning Rate	$1 \times 10^{-5}$
Optimizer	Adam
$\beta_1, \beta_2$	0.9, 0.999
Weight Decay	0.01
Batch Size	1
Loss Function	Cross-Entropy
Logging Steps	100
Evaluation Metric	Accuracy
Global Seed	255

Table 1: Hyperparameters Configuration

## 3.3 Chain of Thought Prompting

We use 6 LLMs for CoT zero-shot: Mixtral(8x7b), Claude, GPT3.5, Llama-2-70b (Touvron et al., 2023), OpenChat (Wang et al., 2023), and Microsoft Copilot<sup>4</sup>. We use graphical web page interfaces for Claude, GPT3.5(chatGPT), and Microsoft

<sup>4</sup>Available at <https://copilot.microsoft.com/>

Copilot (precise mode). We use the same prompt as the ReConcile initial round. Microsoft Copilot gives the best performance of this section.

### 3.4 ReConcile Round Table

Models make mistakes in one or more types of questions and cannot provide the correct answer on the first attempt. We need to ask them to pay attention to certain parts of the question or give hints to the model so it can provide the correct answer. To ensure human involvement is minimized and models can help each other, we have employed the ReConcile method. Using this approach, each model complements the other. The process of this system is as follows:

- **Initial Response Generation:** First, using an initial prompt, we ask each model to provide the answer to the question, provide a reason for the answer, and declare a confidence level between 0 and 1.
- **Multi-Round Discussion:** We give the responses, reasoning, and confidence levels of the three models, along with the initial prompt, as input to the models once again. This enables them to consider both the context of the question and the responses of the three models when making a selection.
- **Final Answer Generation.** In this stage, we initialize a weight for each of the 4 options of the question, and these weights are summed up with the confidence level of each model. Finally, the option with the highest weight is chosen as the correct option.

Let’s denote the confidence of the model  $m_i$  for its selected choice  $c_j$  as  $\text{conf}(m_i, c_j)$ . Then, the total confidence of each choice can be calculated as:

$$TC(c_j) = \sum_{i=1}^3 \text{conf}(m_i, c_j) \quad (2)$$

Where the sum is over all models that selected choice  $c_j$ . Finally, the choice with the highest total confidence is selected as the correct choice:

$$c_{\text{correct}} = \arg \max_{c_j \in C} TC(c_j) \quad (3)$$

This means that the correct choice is the one that maximizes the total confidence over all choices.

The notable point is that this method should be implemented by models that roughly have equal performance to grow together after several rounds. If a model has much lower performance compared to other models, its reasoning and confidence level may negatively affect others. For this reason, we performed this task on three models: Mixtral8x7b, GPT3.5, and Claude, which have almost similar accuracy in the initial round. This iterative process can be continued until all models reach a consensus and all agree on a specific option for the questions. We repeated this process for two discussion rounds.

## 4 Experiments and Results

### 4.1 Experimental Setup

The training and test sets of the sentence and word puzzle datasets are used with a split of 0.8 and 0.2, respectively for fine-tuning. Additionally, we utilize Google Colab’s T4 GPU with the hyperparameters as shown in Table 1. For zero-shot prompting, we use the 120-row test set from the sentence puzzles.

We leverage HuggingChat <sup>5</sup> for Mixtral8x7b, OpenChat, and Llama2-70b. Furthermore, we utilize the official web interface of Claude, Microsoft Copilot, and GPT3.5. Results for each model and their corresponding training codes are available in the GitHub repository.

### 4.2 Results

As illustrated in Table 2 for the sentence puzzles and Table 3 for the word puzzles, we present the performance of BERT and RoBERTa Large in both their base and fine-tuned versions.<sup>6</sup> We load the best model based on Overall Accuracy at the end of each training. The best performance is achieved by RoBERTa for both sentence and word puzzles. As illustrated in Appendix A For every 100 training steps, we log the overall accuracy for two models.

The submission scores computed by the task organizer for CoT zero-shot are available in Table 4. Among these LLMs, Microsoft Copilot achieves the best performance. The success of LLMs in responding to these questions depends on the model’s ability to recognize that these questions are tricky and that it doesn’t need to provide logical reasoning in many cases; the question merely plays with words. Microsoft Copilot understood this phenomenon in many questions. However, it

<sup>5</sup>Available at <https://huggingface.co/chat/>

<sup>6</sup>S(sentence), ori(original), sem(semantic), con(context)

Model	Type	S_ori	S_sem	S_con	S_ori_sem	S_ori_sem_con	S_overall
BERT Base	Baseline	0.400	0.450	0.325	0.350	0.175	0.391
RoBERTa Large	Baseline	0.250	0.175	0.275	0.175	0.050	0.233
BERT Base	Finetune	0.725	0.750	0.650	0.725	0.575	0.708
RoBERTa Large	Finetune	<b>0.800</b>	<b>0.775</b>	<b>0.725</b>	<b>0.775</b>	<b>0.700</b>	<b>0.766</b>

Table 2: Models’ Performance on Sentence Puzzles

Model	Type	W_ori	W_sem	W_con	W_ori_sem	W_ori_sem_con	W_overall
BERT Base	Baseline	0.562	0.343	0.375	0.281	0.093	0.427
RoBERTa Large	Baseline	0.250	0.281	0.343	0.218	0.093	0.291
BERT Base	Finetune	<b>0.687</b>	0.656	0.468	0.625	0.375	0.604
RoBERTa Large	Finetune	<b>0.687</b>	<b>0.687</b>	<b>0.562</b>	<b>0.656</b>	<b>0.468</b>	<b>0.645</b>

Table 3: Models’ Performance on Word Puzzles

also made mistakes in several questions. For example, consider Appendix C for an illustration.

For Reconcile, the results of each model in every round, as well as the consensus reached in each round, are presented in Table 5. In the table, we observe that Claude achieves the highest overall accuracy among the models in each round. Nearly every model in the Reconcile system either improves or maintains its best performance in overall accuracy with each round. This suggests that they are all capable of making informed decisions based on the reasoning provided by all agents during the discussion rounds. At the conclusion of round 2, the consensus overall accuracy stands at 0.758, which is 0.3 to 0.5 points higher than the initial round results of all three models. Furthermore, we note that the consensus result improves by approximately 1 percent from the initial round to round 1, and by approximately 0.8 percent from round 1 to round 2. This indicates that after several rounds, the models converge and reach a consensus on the questions. See Appendix D. Also for an example, see Appendix E

## 5 Conclusion

In this paper, we present our approach to SemEval 2024 Task 9, BRAINTEASER: A Novel Task Defying Common Sense, which challenges models to think creatively beyond conventional reasoning. Through fine-tuning BERT and RoBERTa models, as well as employing zero-shot prompting techniques using various large language models, we achieved notable performance improvements. Particularly, Microsoft Copilot performs the best without being taught beforehand, showing it understands the tricky task really well. Furthermore, our

ReConcile Round Table method demonstrates the efficacy of collaborative decision-making among models, leading to a progressive improvement in overall accuracy across multiple rounds of discussion.

## References

- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. [Reconcile: Round-table conference improves reasoning via consensus among diverse llms](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zi-Yi Dou and Nanyun Peng. 2022. [Zero-shot common-sense question answering with cloze translation and consistency optimization](#).
- Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2024. [Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles](#).
- Yongfeng Huang, Yanyang Li, Yichong Xu, Lin Zhang, Ruyi Gan, Jiaying Zhang, and Liwei Wang. 2023. [MVP-tuning: Multi-view knowledge retrieval with prompt tuning for commonsense reasoning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13417–13432, Toronto, Canada. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris

Model	S_ori	S_sem	S_con	S_ori_sem	S_ori_sem_con	S_overall
Open Chat	0.500	0.500	0.525	0.375	0.300	0.508
Llama-2-70b	0.600	0.625	0.550	0.550	0.400	0.591
Mixtral-8x7b	0.750	0.700	0.650	0.650	0.525	0.700
GPT3.5	0.750	<b>0.775</b>	0.625	0.650	0.500	0.710
Claud	<b>0.775</b>	0.725	0.700	<b>0.725</b>	0.625	<b>0.730</b>
Microsoft Copilot	<b>0.925</b>	<b>0.900</b>	<b>0.775</b>	<b>0.875</b>	<b>0.750</b>	<b>0.860</b>

Table 4: LLM’s Zero-Shot Performance on Sentence Puzzles

Model	S_ori	S_sem	S_con	S_ori_sem	S_ori_sem_con	S_overall
<b>Initial Round</b>						
GPT3.5	0.750	<b>0.775</b>	0.625	0.650	0.500	0.710
Claude	<b>0.775</b>	0.725	<b>0.700</b>	<b>0.725</b>	0.625	<b>0.730</b>
Mixtral-8x7b	0.750	0.700	0.650	0.650	0.525	0.700
Consensus	<b>0.775</b>	<b>0.750</b>	<b>0.700</b>	<b>0.675</b>	0.575	0.740
<b>Round 1</b>						
GPT3.5	0.775	0.725	0.675	0.700	0.570	0.720
Claude	<b>0.800</b>	<b>0.775</b>	<b>0.725</b>	<b>0.725</b>	<b>0.600</b>	<b>0.760</b>
Mixtral-8x7b	0.700	0.725	0.625	0.625	0.450	0.680
Consensus	<b>0.800</b>	<b>0.750</b>	<b>0.700</b>	<b>0.725</b>	<b>0.600</b>	<b>0.750</b>
<b>Round 2</b>						
GPT3.5	<b>0.800</b>	<b>0.800</b>	0.675	<b>0.750</b>	<b>0.600</b>	<b>0.750</b>
Claude	<b>0.800</b>	0.775	0.725	0.725	<b>0.600</b>	<b>0.760</b>
Mixtral-8x7b	0.725	0.725	<b>0.725</b>	0.675	0.550	0.725
Consensus	<b>0.775</b>	<b>0.800</b>	<b>0.700</b>	<b>0.725</b>	<b>0.600</b>	<b>0.758</b>

Table 5: Reconcile Results

- Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th eophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2024a. [Mixtral of experts](#).
- Yifan Jiang, Filip Ilievski, and Kaixin Ma. 2024b. Semeval-2024 task 9: Brainteaser: A novel task defying common sense. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1996–2010, Mexico City, Mexico. Association for Computational Linguistics.
- Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. 2023. [BRAINTEASER: Lateral thinking puzzles for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore. Association for Computational Linguistics.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. [Encouraging divergent thinking in large language models through multi-agent debate](#).
- Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. [Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-

bog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. [Openchat: Advancing open-source language models with mixed-quality data](#). *arXiv preprint arXiv:2309.11235*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#).

Jiarui Zhang, Filip Ilievski, Kaixin Ma, Jonathan Francis, and Alessandro Oltramari. 2022. [An empirical investigation of commonsense self-supervision with knowledge graphs](#).

## A Training logs

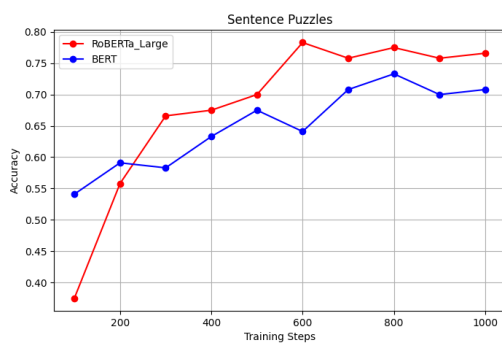


Figure 3: Overall Accuracy of Two Models Logged Every 100 Training Steps on Sentence Puzzles.

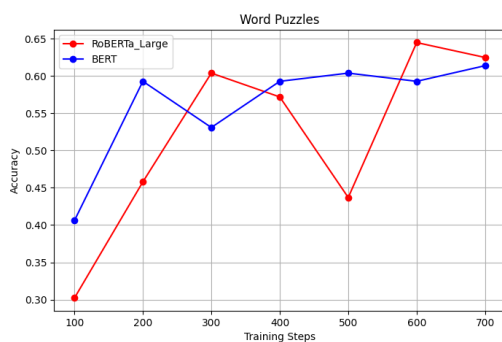


Figure 4: Overall Accuracy of Two Models Logged Every 100 Training Steps on Word Puzzles.

## B Prompts

Initial Prompt

Please answer the question with step-by-step reasoning. Also, evaluate your confidence level (between 0.0 and 1.0) to indicate the possibility of your answer being right. Please give the answer in this template:  
 Solution:  
 Reasoning:  
**{Question}**  
**{Choices}**

Discussion Prompt

**{initial\_prompt}**  
 Carefully review the following solutions from other agents as additional information, and provide your own answer and step-by-step reasoning to the question. Clearly state which point of view you agree or disagree with and why.  
**{agent\_one\_answer}**  
**{agent\_one\_reason}**  
**{agent\_one\_confidence}**  
**{agent\_two\_answer}**  
**{agent\_two\_reason}**  
**{agent\_two\_confidence}**  
**{agent\_three\_answer}**  
**{agent\_three\_reason}**  
**{agent\_three\_confidence}**

Figure 5: ReConcile Initial and Discussion Prompts

## C Copilot Examples

As illustrated in Table 6, the model correctly answers the first three samples but provides an incorrect response for the fourth. This model lacks accuracy in calculation questions, often correctly presenting the equation but solving it inaccurately. Additionally, in some cases, the model accurately formulates the equation and even solves it correctly. However, when the desired option is not among the choices, instead of choosing the 'None of the above' option, it attempts to reinterpret the question in a way that one of the numerical options becomes correct.

## D Reconcile Convergence

Table 7 indicates the number of questions to which all three models provide the same answer choice in each round. This suggests that after each round, the models achieve a relative convergence.

Question	Answer
How can you toss a fragile glass ball against a sturdy brick wall from a distance without shattering it? A. No problem, because sturdy bricks are very hard to crack B. You need to throw the glass ball hard. C. You need to throw the glass ball gently. D. None of above.	Solution: A: No problem, because sturdy bricks are very hard to crack Reasoning: The question is a play on words. It's not the glass ball that we're worried about cracking, it's the brick wall. And brick walls are indeed very hard to crack.
Ana's school has 4 students in Alaska: Maria, James, Jonghyun. What is the name of other student? A. Mark. B. Ana C. Michael D. None of above.	Solution: B: Ana. Reasoning: The question states "Ana's school has 4 students in Alaska: Maria, James, Jonghyun." So, Ana is the fourth student.
In a house, there are three major rooms. The first is stuffed with cash. The second is stuffed with crucial documents. The third is stuffed with jewels. One day, all of these rooms caught fire. Which room did the cops initially put out the fire? A. The one with jewelry. B. The one with important papers. C. The one filled with money. D. None of above.	Solution: D: None of above. Reasoning: The question is a play on words. It's not the cops' job to put out fires, it's the firefighters' job.
During a 5-day competition, a professional fisherman caught 30 fish. He caught three more fish each day than the day before. On the first day, how many fish did the fisherman catch? A. Two B. One C. Three D. None of above	Solution: A: Two. Reasoning: If the fisherman caught 30 fish in 5 days, and each day he caught three more fish than the day before, we can set up the equation: $x + (x+3) + (x+6) + (x+9) + (x+12) = 30$ . Solving for x gives us $x = 2$ . So, the fisherman caught two fish on the first day.

Table 6: Microsoft Copilot Answers to Some Questions

Round	Question Count
Initial Round	74
Round 1	88
Round 2	104

Table 7: ReConcile Models Convergence

## E Reconcile Example

The iterative process of answering the question below is illustrated in Table 8.

**If a boat is parked in a marina and tied to a nearby station, how did its owner found the boat a couple of hours later in the middle of the ocean?**

- A. The station wasn't fixed in the marina
- B. The dolphins untied the knots.
- C. The waves were powerful, which pushed the boat and the marina together.
- D. None of above.



Model	Solution	Reasoning	Confidence
<b>Initial Round</b>			
GPT3.5 Solution	A	If the station (presumably a mooring or dock) wasn't securely fixed in the marina, it could have drifted into the ocean, taking the boat with it.	1.0
Claude Solution	B	This explains how the boat ended up drifting into the ocean after being tied up.	0.7
Mixtral Solution	A	If the station wasn't fixed properly in the marina, it's possible that the boat became loose and drifted away into the ocean due to wind or tidal influences. Over time, the gap between the boat and the dock could increase, eventually pushing the boat further out into deeper waters.	0.85
<b>Round 1</b>			
GPT3.5 Solution	A	If the station to which the boat was tied was not properly fixed in the marina, it could have drifted away into the ocean, taking the boat with it.	0.9
Claude Solution	A	I agree with Agents two and three that an unsecured floating station drifting away best explains how the boat ended up adrift in the ocean.	1.0
Mixtral Solution	A	While all options are plausible scenarios for how the boat might have ended up in the ocean, option A provides the most likely explanation given our current understanding of physics and marine environments. Both agent two and agent three suggest that if the station was not fixed properly, it could have been moved by external forces such as wind or tide, thereby pulling the boat along with it. In contrast, option C assumes that the waves were strong enough to move both the marina and the boat simultaneously, an event that seems less probable than the movement of a single object like the station. Option B attributes the boat's displacement to intelligent behavior on part of dolphins, which lack the necessary appendages to manipulate rope knots underwater. Therefore, we can infer that the station must have been unsecured, allowing it to drift out into the ocean while carrying the boat with it.	0.95
<b>Round 2</b>			
GPT3.5 Solution	A	If the station to which the boat was tied was not properly fixed in the marina, it could have drifted away into the ocean, taking the boat with it.	0.95
Claude Solution	A	I agree with Agents two and three that an unsecured floating station drifting away provides the most logical explanation for how the boat ended up adrift.	1.0
Mixtral Solution	A	All three agents came to the conclusion that option A is the most likely scenario because it's more feasible that an unsecured floating station would be moved by external forces such as wind or tide, rather than assuming that powerful waves managed to push the marina and the boat together at once, or attributing the boat's displacement to intelligent behavior from dolphins who wouldn't possess the physical capability to manipulate rope knots underwater.	0.98

Table 8: Reconcile Example